

Research of Models in the Field of Multimodal Sentiment Analysis

Ruiyang Feng

Department of Computational Science, City University of Hong Kong (Dongguan), Dongguan, China

72400082@cityu-dg.edu.cn

Abstract: As social media is growing, more and more users tend to attach a picture with their posts to convey their tendency. So, it is necessary to find a way to realize the sentiment tendency to label the user, which can act as a reference when the system pushes personal information for the user. This article reviews the articles that make big contributions to Multimodal Sentiment Analysis and compares two types of multiply model sentiment analysis models, Entity-Sensitive Attention Fusion Network (ESAFN) and Multi-Interactive Memory Network (MIMN), from text process, picture process, the structure of the model, the generation of the dataset, and the result. The results indicate that ESAFN outperforms MIMN in both accuracy and MACRO-F1, suggesting that ESAFN demonstrates better stability and reliability across different datasets, potentially due to its more consistent ranking standards and data handling approach. This article also points out some limitations of these two models in model structure and dataset correctly.

Keywords: Multimodal, sentiment analysis, large language model, transformer

1. Introduction

User portrait construction and personalized information push based on big data technology are important means to improve user experience. As an important attribute of user behavior, emotional tendency occupies a significant position in user portraits. Therefore, sentiment analysis is an important and valuable direction of analysis. However, as emotions are unique to humans, it is difficult for machines to predict human emotions in a formulaic way. In addition, with the upgrade of social media and the increase in network bandwidth, users tend to post composite information of text and images. Images carry subtle human emotions, so image information must be considered when conducting sentiment analysis.

Some previous articles make contributions to it. For example, Poria clarifies the baseline of multimodal sentiment analysis, Chen introduces a multimodal sentiment analysis architecture of Gated Recurrent Unit (GRU) with a temporal attention mechanism, which gains 76.5 percent accurate performance [1,2]. Multimodal Sentiment Analysis To Explore the Structure of Emotions suggests a multimodal sentiment aspect analytical method and gets 72% accuracy in the text set [3].

This paper compares two models for social media sentiment analysis using multimodal technology: Entity-Sensitive Attention and Fusion Network (ESANF) and Multi-Interactive Memory Network (MIMN).

2. Method

2.1. ESAFN's Construction

ESANF processes the sentence and picture separately and then combines them for being processed by ESANF [4]. The first step is finding out the preceding text, the following text, and itself of the entity and embedding them to a vector, which can help the model understand the contextual information with multiple entities [5-7]. In the second place, the ResNet, a type of Convolutional Neural Network (CNN), which is used to process pictures, is used to process the picture, being compressed to 214x214 pixels, into a vector [8]. Then, together, the sentence vector and the picture vector in a series of steps to get the result.

2.2. MIMN's Construction

MIMN uses word representation method embedding the aspect into a vector and then using LSTM to bidirectional process it [5,9]. Words and images also are processed by word representation method and LSTM (in Figure 1) (Long Short Term Memory) bidirectional processing to M^T and M^I [9]. Then process the world matrix and aspect vector by multiple layer network to get hidden word h_i and calculate attention weight a_i , finally use weighted average calculation get $V_{text}^{(1)}$. Processing image matrix and aspect vector also get $V_{img}^{(1)}$. Getting $V_{text2text}^{att(i)}$ by using $V_{text}^{(i-1)}$ and $V_{text}^{(i)}$ to search from M^T and getting $V_{img2text}^{att(i)}$ by using $V_{img}^{(i-1)}$ and $V_{img}^{(i)}$ to search from M^T . Then use $\frac{V_{text2text}^{att(i)} + V_{img2text}^{att(i)}}{2}$ to get the $V_{text}^{att(i)}$ of the i-th layer. Also in every layer this model use GRU(Gated Recurrent Unit) (in Figure 1) processing V_{img}^i and V_{text}^i to enhance the model's ability to abstract and retain global memory. In the final layer SoftMax to get the result [10].

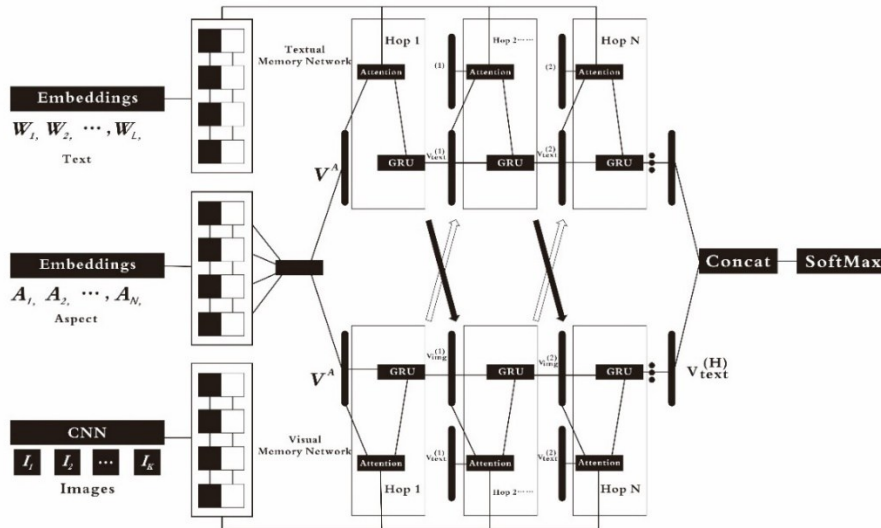


Figure 1: The structure of MIMN [5]

3. Comparison

3.1. Dataset

ESAFN's team creates their dataset from two previous essays' dataset, which contains the user's multiple-model posts in Twitter(X) from 2014 to 2015 and related multiple-model posts from 2016 to 2017 [11,12]. The entities from the dataset cover four classes: Person, Place, Organization, and other. Then invite a specialist in a related field to mark the text and related picture. Cohen's kappa coefficient was used to evaluate the consistency of them, and the label that received the most votes will be considered as the golden label [13]. Also delete data whose label is different. The final data is divided into three parts, which are the train set (60%), development set (20%), test set (20%).

MIMN's team creates their dataset by collecting multiple-model information from a China website ZOL.com. Getting a lot of mobile phone comments and marking them 1 to 6 aspect. Each aspect has a rank 1 to 10 which represents the user's attitude to this aspect.

ESANF's dataset has a larger range of the types of the data, but MIMN's only uses data from ZOL.com in the mobile field. So MIMN's appearance in another field needs further analysis. Also, ESANF's entities evaluation was ranked by a specialist in this field and uses some mathematical method to make sure the data's accuracy, so this dataset is more authoritative, but MIMN's data aspects extraction and aspects marking's approach are not revealed (subjective opinion makes a big influence in marking). The conclusion is that the dataset of ESANF has more accuracy and authority than MIMN's.

3.2. Picture Process

ESANF uses ResNet152 and MINM uses ResNet50 to extract the information in the picture, and ResNet152 has more parameters, so it has a better performance in extracting details in the picture [8,14,15]. ESANF compresses the picture into 224x224 pixels and divides it into 49 son set pictures, which are 32x32 pixels. And then get a matrix that is 2048x7x7, which makes the model more accurately extract details about the entities. MIMN's operation about picture is the same as ESANF's, and the exclusive difference is the ResNet's layer. And MIMN also uses LSTM to bidirectionally process the vector to find the hidden state of the picture.

There is a question in the process picture, that simply compressing the picture will make the picture lose some details, but some pictures' key information is in these details. So simply compressing the picture is not a good way. PCNet maybe a good way to overcome it [16]. Also, there is another problem that there are a lot of posts in social media related to the current online hotspots. In future models, it should add RAG and another related method to make the model realize the newest information [17].

3.3. Result

Due to these two models' use of different datasets and different evaluation way, the result comparison (Tables 1, Table 2) is to be regarded as a reference. The results are as follows:

Table 1: ESANF's appearance in test data

	TWITTER-14	TWITTER-15	TWITTER-17
ACC	72.25	71.36	65.80
MACRO-F1	71.42	64.28	62.00

Table 2: MIMN's appearance in test data

ACC	61.59
MACRO-F1	60.51

It is apparent that the appearance of MIMN is poorer than ESAFN's appearance (Table 1) in accuracy and MACRO-F1, which means the stability and accuracy of MIMN is lower than ESAFN's [18]. The reason is that the data of MIMN may not have an equal standard of ranks, which makes the model can't get the standard and ranking method of the dataset.

4. Discussion

ESANF still does not perform very well when the entity's tendency is normal, and Murahartawaty Arief and Noor Azah Samsudin give a method to identify normal accurately [19]. Also, the structure of MIMN directly connects Inception and LSTM, which leads the model can't be able to find the complex relationship between multiple models through this linear connection. To overcome this, Deng gives a method to help the model understand the complex nonlinear relation [20]. But these mode still have some restrictions and fault. For example, it still needs humans to label sentiment entries, and further analysis can divide the label more specific to guide models to realize the sentiment from more aspects and can finetuning a Large Language Model to extract the sentiment aspect to make the model process complex and various tasks.

5. Conclusion

This paper presents two types of multiply-model semantic analysis model, ESANF and MIMN, and compares them in many aspects, such as picture process, text process, and result. And it is apparent that ESAFN has a border usage and higher usage to face the various posts in social media due to its dataset structure and collection, and the accuracy of ESAFN is higher. MIMN uses GRU and bidirectional LSTM, also giving some inspiration to forward multiply model semantic analysis and saving more training resources. In conclusion, ESAFN and MIMN make a big achievement, which can apply in user picture to accurate push information for user to improve user's experience, in Multimodal but still have some restrictions.

References

- [1] Poria, S. (2018). *Multimodal sentiment analysis: Addressing key issues and setting up the baselines*. *IEEE Intelligent Systems*, 33(6), 17–25.
- [2] Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., Zadeh, A., & Morency, L.-P. (2017). *Multimodal sentiment analysis with word-level fusion and reinforcement learning*. *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017*.
- [3] Hu, A., & Flaxman, S. (2018). *Multimodal sentiment analysis to explore the structure of emotions*. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD 18*.
- [4] Yu, J., Jiang, J., & Xia, R. (2020). *Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 429–439.
- [5] Xu, N., Mao, W., & Chen, G. (2019). *Multi-interactive memory network for aspect based multimodal sentiment analysis*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 371–378.
- [6] Zhang, M., Zhang, Y., & Vo, D.-T. (2016). *Gated neural networks for targeted sentiment analysis*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).
- [7] Liu, J., & Yue. (2017). *Attention modeling for targeted sentiment*. *ACL Anthology*, 572–577.
- [8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- [9] Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). *Analogical reasoning on Chinese morphological and semantic relations*. *arXiv*.

- [10] Cho, K., Van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. arXiv.
- [11] Lu, D., Neves, L., Carvalho, V., Zhang, N., & Ji, H. (2018). *Visual attention model for name tagging in multimodal social media*. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1990–1999.
- [12] Zhang, Q., Fu, J., Liu, X., & Huang, X. (2018). *Adaptive co-attention network for named entity recognition in tweets*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 5674–5681.
- [13] Cohen, J. (1960). *A coefficient of agreement for nominal scales*. *Educational and Psychological Measurement*, 20(1), 37–46.
- [14] Zhang, L., Bian, Y., Jiang, P., & Zhang, F. (2023). *A transfer residual neural network based on ResNet-50 for detection of steel surface defects*. *Applied Sciences*, 13(9), 5260.
- [15] He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition*. arXiv.
- [16] Chen, B., & Zhang, J. (2024). *Practical compact deep compressed sensing*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [17] Ghosh, A., Z., M., & Kalinin, S. V. (2024). *Active deep kernel learning of molecular functionalities: Realizing dynamic structural embeddings*. arXiv.
- [18] Opitz, J., & Burst, S. (2021). *Macro F1 and Macro F1*. arXiv.
- [19] Murahartawaty A., & Noor A. S. (2023). *Neutral class handling for customer sentiment analysis in binary classification: a comparative study of supervised machine learning classification algorithm*, 1–8.
- [20] Deng, L., Liu, B., Li, Z., Ma, J., & Li, H. (2023). *Context-dependent multimodal sentiment analysis based on a complex attention mechanism*. *Electronics*, 12(16), 3516.